ELSEVIER
Available online at www.sciencedirect.com
SciVerse ScienceDirect
Procedia Environmental Sciences 10 (2011) 784 – 789
Procedia
Environmental Sciences

2011 3rd International Conference on Environmental Science and Information Application Technology (ESIAT 2011)

A Functional Measure-Based Framework for Evaluation of Multi-Dimensional Point Access Methods

Mohammadreza keyvnpour[a],Najva izadpanah[b]

[a]*Department of Computer Engineering,Alzahra University,Tehran, Iran*
[b]Department of Computer Engineering, Islamic Azad University, Qazvin Branch,Qazvin, Iran

Abstract

Recent improvement in hardware technology has lead to producing lots of multi-dimensional data such as text, image, graphics and video in various applications. Multi-dimensional access methods have developed for supporting fast retrieval of multi-dimensional data from multi-dimensional databases. In the literature multi-dimensional access methods totally have been classified in to two classes including point access methods and spatial access methods. In this study, we focused on multi-dimensional point access methods, and due to variety of methods have developed in recent decades, we proposed a functional measure-based framework to evaluating multi-dimensional point access methods. In this framework, in order to present a comprehensive evaluation of multi-dimensional point access methods, firstly, we extended related classification of multi-dimensional point access methods in the literatures, secondly, we presented some important functional measures to evaluate multi-dimensional point access methods and we evaluated multi-dimensional point access methods based on these functional measures. We hope this study will lead to development of more efficient methods to support search process in multi-dimensional point data bases.

Selection and/or peer-review under responsibility of Conference ESIAT2011 Organization Committee.

Keywords: Multi-dimensional space; point access methods;point data bases

1. Introduction

Due to the advances in hardware technology and increase in the production of multi-dimensional data such as text, image, graphics and video in various applications, it is necessary to develop efficient methods for supporting fast and efficient search process in such databases.

Multi-dimensional access methods have developed for supporting fast retrieval of multi-dimensional data. Many multidimensional access methods have been proposed during recent decades. In the literature multi-dimensional access methods totally have been classified in to two classes: Point access methods (PAMs) and spatial access methods (SAMs) [1, 2]. PAMs have been designed to perform searches on point databases that store only points that do not have a spatial extension [2]. Points are organized in a number of buckets which in PAMs corresponds to some subspace of the data space. As in [3] PAMs has

Selection and/or peer-review under responsibility of Conference ESIAT2011 Organization Committee.
doi:10.1016/j.proenv.2011.09.127

categorized by the properties of the bucket regions: structures with disjoint regions or structures with mutual overlapping regions, structures that may have the region shape of box or structures that have some arbitrary polyhedral shape regions, structures that cover the complete data space or structures that just cover those parts that contain data points .SAMs have been developed lately to manage extended objects, such as lines, polygons. As in [2] SAMs has classified in to five categories: methods that transforms objects either into a higher dimensional space and then support them by PAMS or in to one dimensional space by Space-Filling Curves and then support by one dimensional structures such as B-tree, structures with overlapping regions. Clipping-based methods that do not allow any overlaps between bucket regions, and Multiple Layers that allow data regions of different layers overlap.

In this study, we focused on multi-dimensional point access methods. Due to variety and plenty of multi-dimensional point access methods that has been proposed in recent decades, we suggested a systematic framework to evaluate multi-dimensional point access methods based on important functional measures. In this framework, in order to present a comprehensive evaluation of multi-dimensional point access methods, firstly, we extended related classification of multi-dimensional point access methods in the literatures and classified these methods; secondly, we presented some important functional measures to evaluate multi-dimensional point access methods and evaluated multi-dimensional point access methods based on these functional measures.

The rest of the paper is organized as follows: we describe multi-dimensional point access methods in section 2. In section 3, the proposed functional measures-based framework for evaluating multi-dimensional point access methods is presented. And, section 4 includes the conclusion.

2. Multidimensional Point Access Methods

Many different multi-dimensional point access methods have been proposed during past decades. The main goal of these structures is to improve the performance in the retrieval of multi-dimensional data points that satisfy a search query. Two most common query types that must be solved by Multi-dimensional point access methods are point query and range query [4, 5].Point query gives a multi-dimensional point and find all data points in the database that overlapping with it. Point query is characterized in (1).

$$P\_Q(q) = \{o \mid q \cap o = q\} \qquad (1)$$

Where, P_Q characterize point query, q represents query point, and o represents data points in the database.

Range query gives a multi-dimensional interval and find all data points in the database that having at least one point in common with that interval. Range query is characterized in (2).

$$R\_Q(q) = \{o \mid I \cap o \neq 0\} \qquad (2$$

Where, P_Q characterize Range query, q represents query point, o represents data points in the database, and I represents multi-dimensional interval.

The points in the database are organized in a number of buckets, each of which corresponds to a disk page and to some subspace of the universe. The structure of multi-dimensional point access methods are either base on hash function or hierarchical partitioning the data space [2].The aim of hash based structures is to construct hashing functions that preserve proximity at least to some extent to minimizing the number of disk accesses per range query [5]. Grid file [6] is the most famous hash based point access method. Most of the multi-dimensional point access methods are based on the principle of hierarchical partitioning of the data space, so that they have a tree-like structure [2, 4].K-d-b-tree [7] is the most famous hierarchical point access method. Fig.1 shows an example of grid file structure and an example of K-d-b-tree structure.

3. A functional Measure-Based Framework for Evaluation of Multi-dimensional Point Access Methods

This section includes the proposed framework for evaluation of multi-dimensional point access methods. Components of this framework include both classification of multi-dimensional point access methods, and evaluation of these methods according to functional measures. firstly, we extend related classification of multi-dimensional point access methods in the literatures and classified these methods; secondly, we presented some important functional measures to evaluate multi-dimensional point access methods and then we evaluated multi-dimensional point access methods based on these functional measures in this section.

3.1. Classification of Multi-dimensional Point Access Methods

According to our study on multi-dimensional point access methods, here, we extend categorization of multi-dimensional point access methods that in [3], and we have presented classification of multi-dimensional point access methods based on their region shape and the manner they partition the data space. Classification of multi-dimensional point access methods is shown in Fig. 2. Complete region refers to this fact that the union of the subspaces is the complete data space. Disjoint or overlapped partitioning refers to this fact that directory regions are either disjointed or they are overlapped. Directory Region shape refers to this fact that the data space is partitioned to the multi-dimensional rectangular or general shape. Fig. 3 shows an example of partitioning the data space in to rectangular versus general shape regions and example of complete versus incomplete partitioning.

3.1.1. Methods with complete partitioning and disjoint rectangular directory regions

In our study, first class of multi-dimensional point access methods is Methods with complete partitioning and disjoint rectangular directory regions .Some examples of these methods are Grid file and K-d-b-tree. For example as it is shown in Fig.1,Grid file partition a data space into a grid structure by splitting each dimension into several non uniformly spaced rectangular regions. A directory is maintained in order to determine the physical location of each grid element. As it is shown in Fig.1, in k-d-b-tree, each internal node corresponds to a rectangular shape region. And the union of regions corresponding to nodes at the same tree level is the complete data space [7].

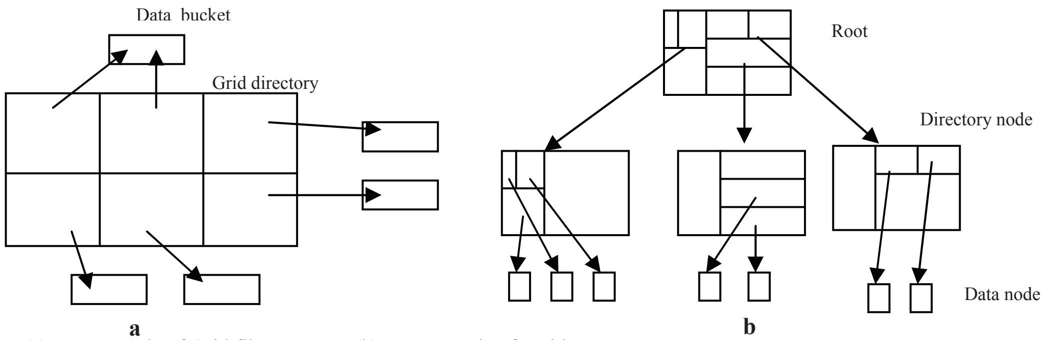

Fig. 1. (a) An example of Grid file structure, (b) An example of K-d-b-tree structure

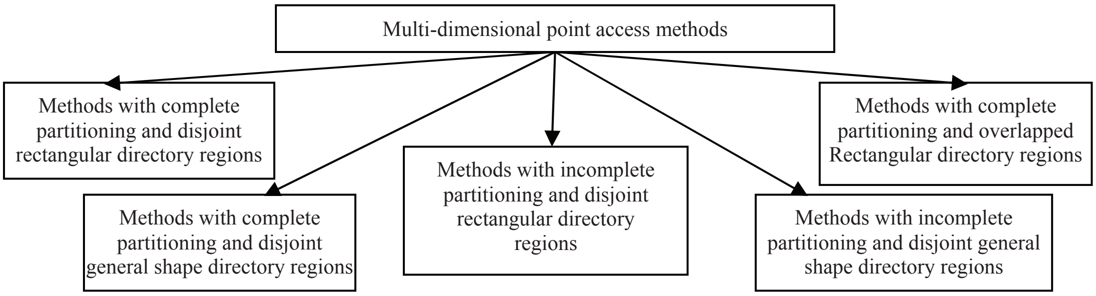

Fig. 2. Classification of multi-dimensional point access methods.

3.1.2. Methods with complete partitioning and disjoint general shape directory Regions

In our study, second class of multi-dimensional point access methods is Methods with complete partitioning and disjoint general shape Regions. Examples of these methods are BSP-tree [8], and HB-tree [9].Partitioning the data space into general shape regions leads to maintain balanced occupancy in split nodes, this lead to increase in storage utilization. For example, BSP-tree is a binary tree that represents a recursive complete partitioning of the data space into subspaces. Each subspace is subdivided independently of its history in to general shape regions. The choice of the partitioning hyper planes depends on the distribution of the data objects in a given subspace [2, 8].

3.1.3. Methods with incomplete partitioning and disjoint rectangular directory regions

In our study, third class of multi-dimensional point access methods is Methods with incomplete partitioning and disjoint rectangular directory Regions. Some examples of these methods are multilevel grid file [10] and buddy tree[3]. The motivation of using incomplete partitioning is because of negative effect of complete partitioning. In complete partitioning of data space, amount of empty space that does not contain the data points may be large. This increases the number of nodes that can intersect the query region, and thus lead to increases the number of disk access during the search. For example, buddy tree is constructed by consecutive insertion. Buddy tree split the data space recursively with iso-oriented hyper planes [3].

3.1.4. Methods with incomplete partitioning and disjoint general shape directory regions

In our study, forth class of multi-dimensional point access methods is Methods with incomplete partitioning and general shape directory regions. An example of these methods is HG- tree [11].HG-tree proposed to balancing the node occupancies by using general shape regions, and reducing the empty space in directory regions by using incomplete partitioning. To achieve the first goal it uses Hilbert curve to apply a linear ordering on the data points and on the directory regions. Hilbert curve is a space-filling curve that mapping the unit interval onto the n-dimensional unit hyper rectangle continuously.

3.1.5. Methods with complete partitioning and overlapped Rectangular directory regions

In our study, fifth class of multi-dimensional point access methods is Methods with complete partitioning and overlapped general shape regions. An example of these methods is Twin grid file [12].The twin grid file tries to increase Space utilization by organizing two dependent grid files at the same time [3]. Both grid files span the whole universe and the distribution of the data points among the two files is performed dynamically. But this fact that data points may be found in either of the two grid files that leads to overlapped directories requires search operations to visit the two files, which causes some overhead [12].

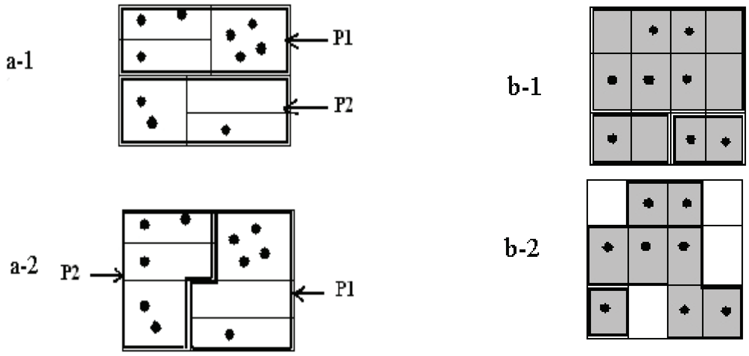

Fig. 3. (a-1) P1 and P2 are rectangular shape regions. (a-2):P1 and P2 are general shape regions. (b-1):Complete partitioning.(b-2):Incomplete partitioning.

3.2. Evaluation of Multi-dimensional Point Access Methods

According to our study on multi-dimensional point access methods, here, we present some important functional measures to evaluate multi-dimensional point access methods according to them.

The functional measures that considered in our evaluation of multi-dimensional point access methods are as follows:1-Space utilization: Space utilization is a measure of how well the available data storage space is used with the access method. An efficient point access method must guarantee high space utilization.2-Disk access during search: as multi-dimensional point access methods save in disk, an efficient point access method must guarantee low Second memory access during search. Because disk device is slow, and high number of access to disk during search reduces the speed of retrieval process.3-Support for non uniform data distributions: by increasing the dimensionality of datasets they tend to have non uniform distribution, it is important for multi-dimensional point access method to preserve their efficiency for non uniform data distributions.

In Methods with complete partitioning and disjoint rectangular regions because of using rectangular directory regions, amount of empty space that does not contain the data points may be large This increases the number of nodes that can intersect the query region, and thus lead to increases the number of disk access during the search. And if the data distribution is non uniform, it is clear that large parts of the space are empty, and then leads to very poor space utilization. Methods with complete partitioning and disjoint general shape directory regions Because of using general shape directory regions that leads to maintain balanced occupancy in split nodes, and amount of empty space that does not contain the data points decrease this lead to better space utilization especially for non uniform data distributions. Methods with incomplete partitioning and disjoint rectangular directory regions because of using incomplete partitioning leads to decrease amount of empty space that does not contain the data points and thus leads to decrease the number of disk access during the search. Methods with incomplete partitioning and disjoint general shape directory regions because of using incomplete partitioning and general shape directory regions have good performance in case of space utilization and disk access during search for uniform and not uniform data distributions. Methods with complete partitioning and overlapped Rectangular directory regions try to increase Space utilization by using overlapped directory region, thus they have better. But this fact that data points may be found in one or the other overlapped regions that leads to overlapped directories requires search operations to visit all overlapped regions, which causes to increase in disk access search. And because of using complete partitioning and rectangular directory region these structures can’t support non uniform data sets. Evaluation of multi-dimensional point access methods is summarized in Tab 1.

4. Conclusion

Due to importance and variety of multi-dimensional point access methods, in this study, we proposed evaluating multi-dimensional point access methods in a functional measure-based framework. The proposed frame work in this study includes two components: classification of multi-dimensional point access methods, and evaluation of multi-dimensional point access methods. Study on multi-dimensional point access methods shows that multi-dimensional point access methods can classified in to four main classes. According to our study on multi-dimensional point access methods, here, we extend categorization of multi-dimensional point access methods that in the literatures, and we have presented classification of multi-dimensional point access methods based on their region shape and the manner they partition the data space in to five classes. Then in this study we proposed some functional measures to evaluate multi-dimensional point access methods. And then we evaluated multi-dimensional point access methods according to functional measures.

Table 1. Evaluation of multi-dimensional point access methods

Multi-dimensional point access methods	Functional measures		
	Space utilization	Disk access during search	Support for non uniform data distributions
Methods with complete partitioning and disjoint rectangular regions	poor	Poor	poor
Methods with complete partitioning and disjoint general shape directory regions	medium	poor	good
Methods with disjoint incomplete partitioning and rectangular directory regions	good	medium	poor
Methods with incomplete partitioning and general shape regions	good	good	good
Methods with complete partitioning and overlapped Rectangular directory regions	medium	Poor	poor

References

[1]K. Markov, K. Ivanova, I.Mitov, and S. Karastanev, " Advance of the Access Methods," International Journal of Information Technologies and Knowledge, Vol.2, 2008.

[2] V. Gaede and O. Günther, " Multi-dimensional Access Methods,"ACM Computing Surveys, Vol. 30, No. 2, 1998.

[3]B.Seeger, B. and H.-P. Kriegel, "The buddy-tree: An efficient and robust access method for spatial data base systems", In Proc 16th Int. Conf. on Very Large Data Bases,1990, p. 590-601.

[4]C.Bohm, S. Berchtold, and D. A. KEIM, "Searching in High-Dimensional Spaces Index Structure for Improving the performance of Multimedia Databases," ACM Computing surveys, vol.33, No.3, ,2001.

[5]H.Samet , Applications of Spatial Data Structures. Reading, MA: Addison-wesley.

[6]J.Nievergelt, H. Hinterberger, and K. C. Sevcik, "The grid file: An adaptable, symmetric multi key file structure", ACM Trans. Database Systems 9 (1), 1984,p.38-71.

[7]J.T. Robinson. "The K-D-B-tree: A search structure for large multidimensionaldynamic indexes". In Proc ACM SIGMOD International Conference on Management of Data, 1981, pp. 10-18.

[8]H.Fuchs, Z.Kedem, and B.Naylor," On visible surface generation by a priori tree structures", Computer Graphics 14 (3),1980.

[9]D.B.Lomet,B.Salzberg,"The hB-tree:A robust multiattribute search structure,Proc.fifth Int,1989.

[10]K.-Y .Whang, and R. Krishnamurthy ,"Multilevel grid files", Yorktown Heights, NY: IBM Research Laboratory,1985.

[11]Guang-Ho Cha and Chin-Wan Chung," HG-Tree: An Index Structure for Multimedia Databases," Proc. IEEE multimedia ,1996.

[12]A.Hutesz, H.-W. Six and P. Widmayer, "Twin grid files: Space optimizing access schemes", In Proc. ACM SIGMOD Int. Conf. on Management of Data,, p.183-190.